

# 2017 July 2 周报

Junhua Lu

## Done

1. Back to home and went through some formalities.
2. Delete useless logs (empty and unmatched)/ remove redundancy/ statistics about different categories/ timespan inspection and statistics of movie data for wiki data
3. Learn high-order function / OO JavaScript syntax, with ES6 features

## To Do

1. Revision of BeXplorer, and help with one student in other processing the wiki data & model building

## 论文阅读:

### WWW2016 *Disinformation on the Web: Impact, Characteristics and Detection of Wikipedia hoaxes*

Hoax 是网络上那些故意移花接木制造虚假信息混淆视听的做法. 本文作者做了不少互联网上反社会行为检测方面的工作, 从数据出发发现特征\再结合一些经验或者前人的知识构建分类器做预测. 虽然给人感觉有点特例但是这些问题都还是有直接的社会问题, 具有一定价值. 本文对于这种 hoax 行为, 是从四方面考虑的: 外表(缺少文字\外链) 文章间的网络结构 其余文章指向它的情况 以及本身创建者的特性. 其采用简单分类器, 效果会比众包的人力检测结果要好. 可视化, 仍然可以作为一种介入分析\构建分类器\特征选择的工具.

**KDD 15** *VEWS: A Wikipedia Vandal Early Warning System*: Vandal 行为 维基百科上的野蛮编辑行为: 无价值的\冒犯性的\破坏性移除内容. 目标, 在尽量少的编辑中检测(预警)vandals. 用的数据是 34000 个编辑者(一半是 vandal), 77 万次编辑(16000 次是被 vandals 编辑的), 时间跨度 2013.1~2014.7. 这里从两方面出发: 编辑行为以及编辑行为的状态转移(在序列中的模式). 对于后者, 由于序列转移的话每个人不一定是相同的编辑数目, 于是为了让特征向量等长, 采用了 autoencoder 这个神奇的东西, 一种基于神经网络的压缩编码方式(把转移矩阵压缩为等长向量, 再拿去分类).

### EUROVIS 17 *Cycle Plot Revisited: Multivariate Outlier Detection Using a Distance-Based Abstraction*

对 cycle plot 的一种修改, cycle plot 平时不怎么说, 可以简单理解为周期性的一维时序数据的折线图, 对于有一定周期性的数据有较好的展示效果, 有时候也可以展现极值和异常值. 但是对于多维数据就做不到了. 为了实现多维的展现, 并且仍然兼顾一些离群值探测的工作. 首先对于周期性数据分组(比如一个月内周一一组,周二一组,)然后根据样本计算组内的某个指标(均值中值), 在计算整体的指标, 并用马氏距离来计算他们的偏移程度. 马氏距离的计算, 对于离群检测有比较大的影响, 但作者这种距离本来在统计上就适合做离群检测, 只要选择合适的距离度量的值(在马氏距离中间的那个矩阵). 系统可视编码和之前看的一样都很简洁, 但是这个马氏距离似乎是对文章添彩了.

看了一些网络表示学习简介: 设定优化目标, 给他们做一个向量的表示. 其实这有点像上面的 autoencoder, 也像现在流行的 X2vec. 好几种方法仍然是借鉴了 word2vec 的思想, 因为在 youtube 的社交图上节点随机游走的出现频次分布和 wikipedia 上文章词的分布有类似之处. 一点网络能够向量化表示, 后面能做的事情就多了. 对于以前一些社交网络问题, 也相当于是有了新的路子.

**DeepWalk:** 借鉴了 word2vec 的思想, 对图一个节点出发, 向周围采样直到最大路径长度, 然后用随机梯度下降学向量.

**LINE WWW2015** 对点之间关系, 考虑两种 一阶 proximity(直接连接, 有强联系)和二阶 proximity(有很多共同链接, 说明也有关系). 在此基础上构建函数去优化, 求得表示向量.